This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).

Author Names & Affiliations

- Colin Phillips - University of Maryland

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

linguistics, language technology, language learning

Title of Submission

Integrating Global Language Resources

Abstract (maximum ~200 words).

Global advances in language technology, education, health, commerce, and disaster response depend on the availability of interoperable resources in thousands of languages. But current knowledge is distributed, uneven, and rarely interoperable. Around 6000 languages spoken in the world today. Extensive, well organized digital resources are available in less than 1% of these languages, and those that exist are often designed for specific audiences or user needs. There is a need for integrated resources that are publicly available, readily extensible via user input, interoperable, and comprehensible to users with diverse goals and levels of expertise. In order for a distributed global network of experts to be motivated to contribute, it is also essential that the resources or standards have a long-term trajectory.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Resources for research on the world’s 6000+ languages are highly uneven, and rarely integrated. Languages like English, Chinese, French, or Arabic are outliers, with extensive corpora of texts, sound recordings, grammars, geographical data, etc. The vast majority of the world’s languages are ‘low resource’ languages, even languages spoken by tens of millions of people. In many cases, the available resources are useful only for specific users or research applications.

Broad language resources have great value for many different applications in fundamental science and applied research. This includes multilingual natural language technology (machine translation, text summarization, speaker identification, etc.), and the need to be able to work with diverse dialects. It includes research on linguistics and language diversity, seeking to understand the fundamental building blocks of human language, and the ways in which language is used to convey social information in diverse settings. It is important for first and second language acquisition, both for understanding how naturalistic learning occurs and for understanding how to improve instructed learning, especially in less commonly taught languages. It includes interventions to promote that development of speech, language, literacy, and higher intellectual skills across diverse populations. And it includes applications for disaster relief, diplomacy, security, and commerce. To take a scenario that occurs frequently: if an earthquake, disease outbreak, or security threat occurs in a location where very little is known about the local languages, how can we create tools and information that can help people on the ground as quickly as possible? Addressing challenges like this requires research that draws on expertise in multiple fields.

Existing expertise is highly distributed, and also of uneven quality. For some languages excellent resources are available, but they are all written in Russian. For many languages some audio recordings are available, thanks to extensive missionary work, but the recordings may be inauthentic because they are recordings of non-native speakers. In some languages the resources are available only in a format that is useful to technologists, or language typologists. And in most cases there is simply very little information available digitally. In those cases, the expertise simply resides in the speakers themselves.

It is not feasible to turn all languages into high-resource languages like English or Chinese. So the greatest progress will come from learning how to do more with less, e.g., can we create language technologies that learn from 1 million words of text what is currently learned from 20 million words of text. And it will also come from learning how to get a head-start on under-resourced languages by leveraging what we already know about closely related well-resourced languages. This approach holds a lot of promise, but it presupposes a good understanding of relatedness among languages.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

In order to address these challenges, we need integrated, interoperable resources, with the scope for integrating new material via expert contributions and via crowd-sourcing. The resources need to be accessible through multiple languages, though this could rely on lingua francas, e.g., a Swahili interface could be useful for engaging with speakers of many East African languages. Also, the resources need to be accessible to users with diverse interests and levels of expertise.

Some existing examples of broad language resources are useful reference points.

1. Typological/comparative databases, e.g., World Atlas of Language Structures (WALS), AutoTyp (NSF 96-16448) and Syntactic Structures of the World’s Languages (SSWL: NSF BCS-0817202). These projects collect properties of languages (e.g., verb-finality, ±uvular consonants) at a specific grain of analysis, aimed at a specific research community. They cover samples from dozens to hundreds of languages. The public sites vary in their technical implementation, accessibility, and future trajectory.

2. Ethnologue. This compendium of language information is uniquely broad in its coverage, including practically all known languages. The focus is on demographics, language families, and vitality, and the maps are static images, not linked resources. Ethnologue recently introduced a pay-for-access model, limiting opportunities for integration with other language resources.

3. DOBES, The Language Archive (Max Planck Inst.), Endangered Languages Project (Google & NSF BCS-1058096 / BCS-1057725). These are collections of media, wordlists, corpora, and smaller analytic samples, primarily for endangered languages. The projects have valuable resources, and special expertise in curating and protecting information from minority languages. The future of some archives is uncertain.

4. OLAC. OLAC allows users to find resources that may be located in many different databases and archives. The goal of linking different repositories to increase their discoverability is a hugely important one. OLAC effectively catalogs a vast array of language resources.

5. Linguistic Data Consortium. LDC is a successful and sustainable consortium that has existed since 1992. LDC provides quality language resources for the language technology community, but at costs that preclude broader engagement. Also, for contributors LDC’s model creates a high bar to entry, making it less attractive for small/individual providers.

6. Langscape. (Phillips is the PI of this project, based at the University of Maryland.) Langscape is an entirely open resource that aims to

aggregate information on the world’s languages through an easy to use GIS interface. The focus is on gathering expertise from experts around the world, and making it available to users with different levels of expertise and applications. It also has a non-public mirror version that is usable on government computing systems and that can integrate closed information sources. The mirror is maintained by a different group. The existing release is a proof of concept for the GIS interface. Future development plans include hosting diverse map layers, developing data standards for interoperability via a WikiData initiative, and developing crowdsourcing capabilities.

In summary, many promising components of a solution exist. But it is important to create possibilities for funding further development.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

It is essential that the language resources be sustainable. Experts will not make the effort to contribute their knowledge if they don’t see the resource as a long-term project.

Also, experts have differing motivations for contributing their knowledge. Some are interest in broad public visibility. Some are only interested in reaching academic audiences. Some are most interested in serving local language communities or diaspora communities.

Consent Statement

- “I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).”